supplementary: When Visual State Space Model Meets Backdoor Attacks

Sankalp Nagaonkar
IIT Dharwad
210020031@iitdh.ac.in

Achyut Mani Tripathi
IIT Dharwad
t.achyut@iitdh.ac.in

Ashish Mishra
HPE lab, Bangalore
mishraashish632@gmail.com

1. Ablation Analysis

- To demonstrate the effectiveness of our proposed approach beyond image datasets, we conducted experiments on the EPIC audio dataset, as shown in Table 6. The results indicate that our backdoor attack methods are effective not only on image datasets but also on audio datasets.
- We also tested various versions of the VMamba model, including Mqamba-in-Mamba (Mim) [1] and Eff-Mamba [2], across both the CIFAR-10 and ImageNet-1K datasets. These experiments confirm that our proposed backdoor attacks are effective across different Mamba model variants. The results are detailed in Table 2 for CIFAR-10 and Table 3 for ImageNet-1K.
- Additionally, we conducted an ablation study to assess the impact of varying the number of swapped rows and columns in our approach. We experimented with 80, 100, and 150 swapped rows and columns, with results presented in Table 5.
- Furthermore, we evaluated the efficacy of the proposed method under All-to-One attacks for the ImageNet-1K dataset. The results are provided in Table 1.
- Table 7 illustrates the creation time of backdoored images for different attacks. The BadNets attack requires the shortest time, while the WaNet attack demands the most time. The time required by the proposed four attacks is comparable to that of the R-Fool attack.
- To better understand the impact of the attack on clean images, we extracted and visualized the layer-wise attention maps for both clean and attacked images in Figures 1 and 2, respectively. We observed that, for clean images, the model focuses on relevant features associated with the class. In contrast, for attacked images, the model tends to focus on irrelevant parts, which misleads the model.

Dataset	Source/Target Pair	Attack	Model	PMA	ASR
ImageNet-1K	Source: "Warplane", "Computer", "Bedroom" Target: Speedboat	S-QRDBA: Both Swap-100	ResNet-18	79.8	78
			ResNet-50	79	89.8
			MLP-mixer	71.3	60
			ViT	75	71.22
			VMamba	68	58
			Mamba in Mamba	89	78
			Efficient Mamba	88.7	76.2

Table 1. Performance Metrics for Different Models on Imagenet with Various Source/Target Pairs(All-to-One setup)

Source/Target	α	Model	Dataset	PMA	ASR
class 3/class 6	0.1	Mim [1]	cifar10	67.3	33.4
		Eff-Mamba [2]		68	54
	0.15	Mim		60	70
		Eff-Mamba [2]		63	44
	0.2	Mim		66	62.6
		Eff-Mamba [2]		72.8	62.8

Table 2. CMA and ASR for Different version of Vmamba Models and α Values on CIFAR-10 Using W-QRDBA Attack

Source/Target	α	Dataset	Model	PMA	ASR
warplane/speedboat	0.1	Mim	Imagenet	84	30.4
		Eff-Mamba [2]		83	35
	0.15	Mim		82.5	50
		Eff-Mamba [2]		80	40
	0.2	Mim		80	69
		Eff-Mamba [2]		77.3	60

Table 3. CMA and ASR for Different version of Vmamba Models and α Values on Imagenet Using W-QRDBA Attack

Source/Target	Attack	Model	Dataset	PMA	ASR
warplane/speedboat	BadNets	Mim	Imagenet	84.5	27
		Eff-Mamba [2]		78.4	35
	R-Fool	Mim		84.3	50
		Eff-Mamba [2]		78	40
	WaNet	Mim		82	37
		Eff-Mamba [2]		77	35

Table 4. CMA and ASR for Different Attacks and Models on Imagenet-1K

Source/Target	No. of Swapped Rows	Model	Dataset	CMA	PMA	ASR
Warplane/Speedboat	80	ResNet-18	Imagenet-1K	90.53	94	68.8
		ResNet-50		88	87	64
		ViT		77.8	77	46.8
		VMamba		86.94	74	40
		MLP-mixer		99.47	75	38.8
	100	ResNet-18		91.88	90	74.8
		ResNet-50		88.94	90	83.8
		ViT		73.94	70	54
		VMamba		86.7	75	49.2
		MLP-mixer		99.8	87	40
	150	ResNet-18		91	90	96.2
		ResNet-50		91.65	88	96.2
		ViT		85.65	76	78.8
		VMamba		89.61	76	40
		MLP-mixer		100	87	80.8

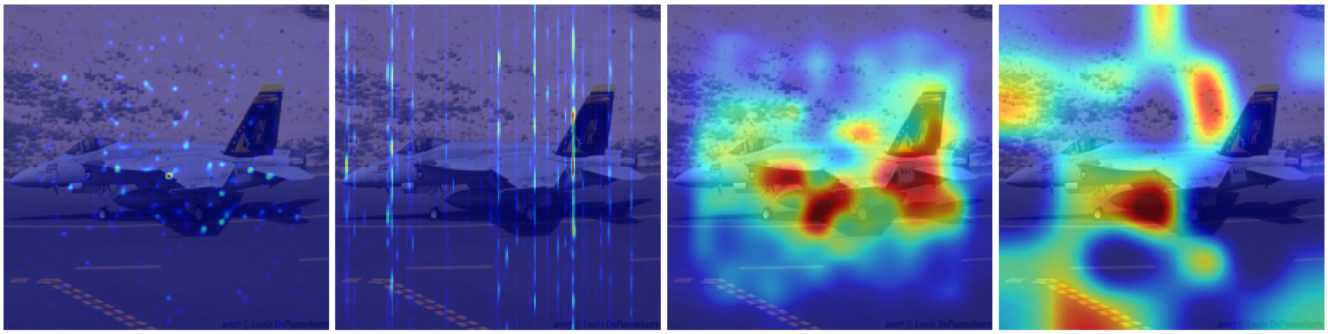
Table 5. Performance Metrics for Several Models for Different Number of Swapped Rows on Imagenet-1K Using S-QRDBA (Both) Attack

Dataset	Source/Target Pair	Attack	Model	PMA	ASR
EPIC	”Scrub, Scrape, Scour, Wipe / Tap Opening, Water”	Badnet	ResNet-18	62.1	61.2
			ResNet-50	62.5	72.7
			MLP-mixer	75.7	94.5
			ViT	78.6	87.1
			VMamba	73.7	70
			MiM	78.1	28.5
			EF-Mamba	73.6	71
		WaNet	ResNet-18	72.2	63.9
			ResNet-50	60.1	77.1
			MLP-mixer	78.2	81.3
			ViT	77.6	55.8
			VMamba	70	50
			MiM	72.4	38.8
			EF-Mamba	70	40
		R-Fool	ResNet-18	60	71.5
			ResNet-50	55.4	91.7
			MLP-mixer	77.8	93.6
			ViT	75.5	80.7
			VMamba	71.5	60.2
			MiM	80.3	66
			EF-Mamba	72	79.4
		S-QRDBA (both swap): 100	ResNet-18	68	90
			ResNet-50	64	95
			MLP-mixer	75	96
			ViT	73.2	92
			VMamba	75	96
			MiM	80	93
			EF-Mamba	84.5	95

Table 6. Performance of The Proposed Attack on Audio Dataset

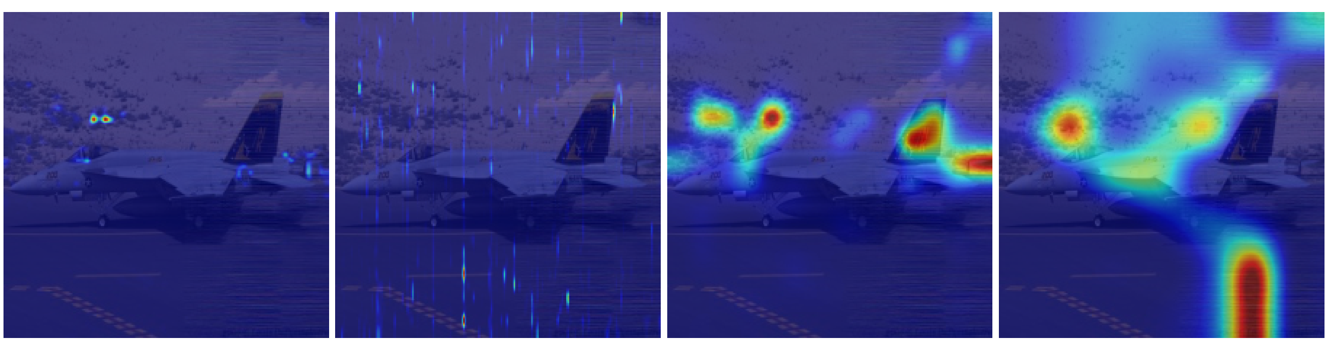
Attacks	Variation	Time (Seconds)
S-QRDBA (Both)	80 Rows, 80 Columns	0.024
	100 Rows, 100 Columns	0.028
	150 Rows, 150 Columns	0.0284
S-QRDBA (Row Swap)	80 Row	0.0227
	100 Rows	0.0241
	150 Rows	0.0259
S-QRDBA (Column Swap)	80 columns	0.0226
	100 Columns	0.025
	150 Columns	0.0263
W-QRDBA	$\alpha = 0.1$	0.0253
	$\alpha = 0.15$	0.025
	α =0.2	0.027
Wanet	-	0.5325
R-Fool	-	0.01187
BadNets	-	0.0000057

Table 7. Table with running time analysis



(a) Layer-1 (b) Layer-2 (c) Layer-3 (d) Layer-4

Figure 1. Layer-wise attention map visualization for cleaned image



(a) Layer-1 (b) Layer-2 (c) Layer-3 (d) Layer-4

Figure 2. Layer-wise attention map visualization for attacked image

References

[1] Tianxiang Chen, Zhentao Tan, Tao Gong, Qi Chu, Yue Wu, Bin Liu, Jieping Ye, and Nenghai Yu. Mim-istd: Mamba-in-mamba for efficient infrared small target detection. *arXiv preprint arXiv:2403.02148*, 2024. 1, 2

[2] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024. 1, 2